Programs for constructing phylogenetic trees and networks of closely related sequences

Naruya SAITOU

Laboratory of Evolutionary Genetics
National Institute of Genetics
Mishima, 411-8540 Japan

Nowadays increasingly many closely related sequences are deposited to the DDBJ/EMBL/GenBank nucleotide sequence database. To deal with those vast number of closely related nucleotide sequences, I recently developed a series of programs to process closely related mass sequence data. We start from BLAST homology search. BLAST, developed by Altschul and others (1990), is usually used for finding sequences that are homologous to the query sequence from target database, but this program is also useful to retrieve closely related sequences. Figure 1 shows example of BLAST output. Human mitochondrial DNA D-loop sequence (185bp) was used for query. A simple program (p0) then extract sequences homologous to query sequence and produce FASTA format output file (see Figure 2). Figure 2 shows only first 6 retrieved sequences.

After retrieving those homologous sequences, mul-

```
>D84917|D84917  Human mitochondrial DNA for D-loop.
  Length = 483

  Plus Strand HSPs:

 Score = 925 (255.6 bits), Expect = 1.9e-71, P = 1.9e-71
 Identities = 185/185 (100%), Positives = 185/185 (100%), Strand = Plus / Plus

Query:    1  ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   67  ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 126

Query:   61  AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  127  AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 186

Query:  121  ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  187  ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 246

Query:  181  ACCCC 185
             |||||
Sbjct:  247  ACCCC 251
```

Fig. 1. BLAST output example.

```
>D84917|D84917  Human mitochondrial DNA for D-loop.
      67   ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 126
     127   AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 186
     187   ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 246
     247   ACCCC 251
>D84909|D84909  Human mitochondrial DNA for D-loop.
      66   ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 125
     126   AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 185
     186   ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 245
     246   ACCCC 250
>D84920|D84920  Human mitochondrial DNA for D-loop.
      66   ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 125
     126   AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 185
     186   ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 245
     246   ACCCC 250
>AA075567|AA075567  zm88f07.s1 Stratagene ovarian cancer (#937219) Homo sapiens
      68   GGGGTCATCCATGGGGACGAGAAGGGATTTGACTGTAATGTGCTATGTACGGTAAATGGC 127
     128   TTTATGTACTATGTACTGTTAAGGGTGGGTAGGTTTGTTGGTATCCTAGTGGGTGAGGGG 187
     188   TGGCTTTGGAGTTGCAGTTGATGTGTGATAGTTGAGGGTTGATTGCTGTACTTGCTTGTA 247
     248   AGCAT 252
>M58068|HUMMTDLR12  Human mitochondrial D-loop region.
     171   ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 230
     231   AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 290
     291   ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 350
     351   ACCCC 355
>M58074|HUMMTDLR18  Human mitochondrial D-loop region.
     171   ATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAA 230
     231   AGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTAC 290
     291   ATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATG 350
     351   ACCCC 355
```

Fig. 2. FASTA format example.

tiple alignment usually follows. However, when we restrict our search only to closely related sequences, multiple alignment is not necessary, for BLAST already extracted homologous regions and those rarely have gaps. Therefore, we can skip multiple alignment process which often takes a very long computer time compared to BLAST search. Therefore, FASTA format output file such as Figure 2 can be easily transformed to multiple-aligned sequence format, as shown in Figure 3. This is output file 1 of program p3, and it consists of multiple alignment part and sequence name part. Figure 3 shows only end of multiple alignment part and start of sequence name part of a big output for 1,516 human mitochondrial DNA D-loop region. Plus and minus sign after last sequence (ID number = 1516) designate variant and invariant nucleotide sites, respectively. Program p3 eliminates invariant sites after producing output 1 and produce output 2 which includes only variant sites. When we deal with closely related sequences, many invariant site are expected to exist, and this procedure can reduce the data file extensively.

Program p4 then examine sequence identity, and all the identical sequences are joined. Consequently, only different sequences remain in output file 2 of program p4 (figure 4). In this example, only 742 mutually different sequences are extracted out of 1,516 individual sequences. Therefore, sequence with new ID 2 is identical to sequence with old ID 125. This means sequences 1 - 124 were all identical.

```
1507.............................t.....g................................................t...........t................
1508.............................t.....g................................................t............................
1509.............................t.........................c............................t..............tt............
1510...................a.........t....................................t...t.........a...t..............g.............
1511...................a.........t....................................t...t.....t.......t..............g.............
1512.............................t....c.................................................t...............t............
1513.............................t........................................t.....t.......t............................
1514.............................t.....g............c...................................................t............
1515.............................t............................................................g..tc..................
1516...............c.............t...............................a...................................................-
    --------++--+-+--+++++++++++++--+-++++++++-++++++++-+++--+-++++++++-++++-++++-++--++----++-++++++++++++++++++++-

seq id  sequence name
     1  D84917|D84917  Human mitochondrial DNA for D-loop.
     2  D84909|D84909  Human mitochondrial DNA for D-loop.
     3  D84920|D84920  Human mitochondrial DNA for D-loop.
     5  M58068|HUMMTDLR12  Human mitochondrial D-loop region.
     6  M58074|HUMMTDLR18  Human mitochondrial D-loop region.
     7  M58075|HUMMTDLR19  Human mitochondrial D-loop region.
     8  M58087|HUMMTDLR31  Human mitochondrial D-loop region.
     9  M58090|HUMMTDLR34  Human mitochondrial D-loop region.
    10  M58094|HUMMTDLR38  Human mitochondrial D-loop region.
    11  M58096|HUMMTDLR40  Human mitochondrial D-loop region.
    12  M58098|HUMMTDLR42  Human mitochondrial D-loop region.
    13  M58100|HUMMTDLR44  Human mitochondrial D-loop region.
    14  M58101|HUMMTDLR45  Human mitochondrial D-loop region.
    15  M58114|HUMMTDLR58  Human mitochondrial D-loop region.
    16  M58139|HUMMTDLR83  Human mitochondrial D-loop region.
    17  X73302|MIHSMUMA   H.sapiens mitochondrial control region DNA, (Berriac).
    18  U33386|HSU33386   Human mitochondrial control region I, sample 3.03 Mongolia.
    19  U33384|HSU33384   Human mitochondrial control region I, sample 3.01 Mongolia.
    20  U59021|HSU59021   Human mitochondrial control region I, sample TUK20 Turkey.
    21  U59022|HSU59022   Human mitochondrial control region I, sample TUK22 Turkey.
```

Fig. 3. p3.out1 file example.

The next step is to eliminate 'single polymorphic' sites in which only one sequence have different character and all others have the same character. This is conducted by program p5. In this fashion, we can rapidly extract so-called phylogenetically informative sites for maximum parsimony methods. However, there may exist bunch of equally parsimonious trees when there are many parallel substitutions, such as transitions in mitochondrial DNA. Therefore, network, generalization of tree, seems to be more suitable for delineating sequence information, as clearly shown by Bandelt and others (1995). Saitou and Yamamoto (1997) also constructed phylogenetic networks for ABO blood group gene sequences and showed its usefulness. Figure 5 illustrates difference between tree and network. When there are incompatible sites in terms of nucleotide configuration (such as sites 6-7 and 8), a rectangular structure emerges.

Figure 6 is example of phylogenetic network for 52 human mitochondrial DNA sequences. There are so many rectangles in this network, and those were found after processing sequence data by using programs described above. We already applied these programs for analysis of human mitochondrial DNA sequences (Oota and others, 1999). Executable prototype programs both for Macintosh and Windows are available upon request to me (email: nsaitou@genes.nig.ac.jp).

```
SSJ aligned format
742 143 3

nucleotides 1 - 50
   1    1 agataagcaatcaaccctatatcacactcaactgcactcaagccaccctc
   2  125 ................t.................................
   3  126 ..................................................
   4  127 ..........c.......................................
   5  128 ..................................................
   6  130 ..................................................
   7  133 ..................................................
   8  139 ..................................................
   9  143 ..................................................
  10  144 ..................................................
  11  145 ..................................................
  12  147 ..................................................
  13  150 ...........................................t......
  14  151 .....................................c............
  15  152 ..................................................
  16  153 ..................................................
  17  154 ..................................................
  18  157 ..................................................
  19  160 ...............................................t..
  20  163 .....g............................................
  21  165 ..................................................
  22  167 ................................................a.
  23  171 ................................c.................
  24  172 ..................................................
  25  184 ..................................................
  26  187 ..................................................
  27  188 ..................................................
  28  201 ................................................c.
  29  209 ..................................................
  30  221 ...............t..................................
```

Fig. 4. p4.out2 file example.

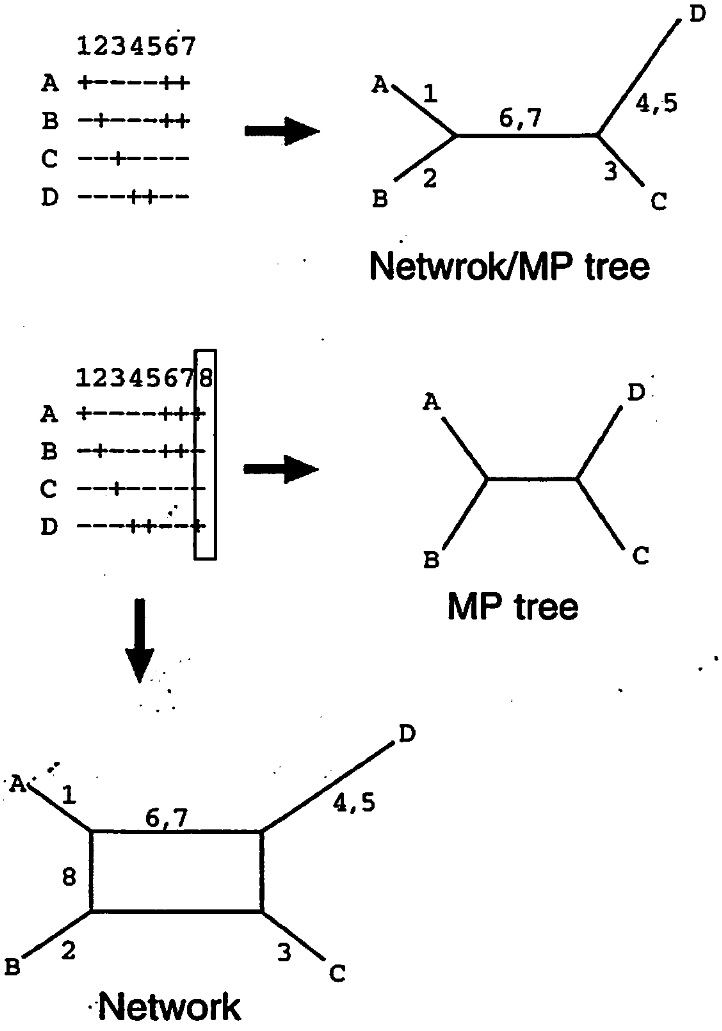

Fig. 5. Network as generalization of tree.

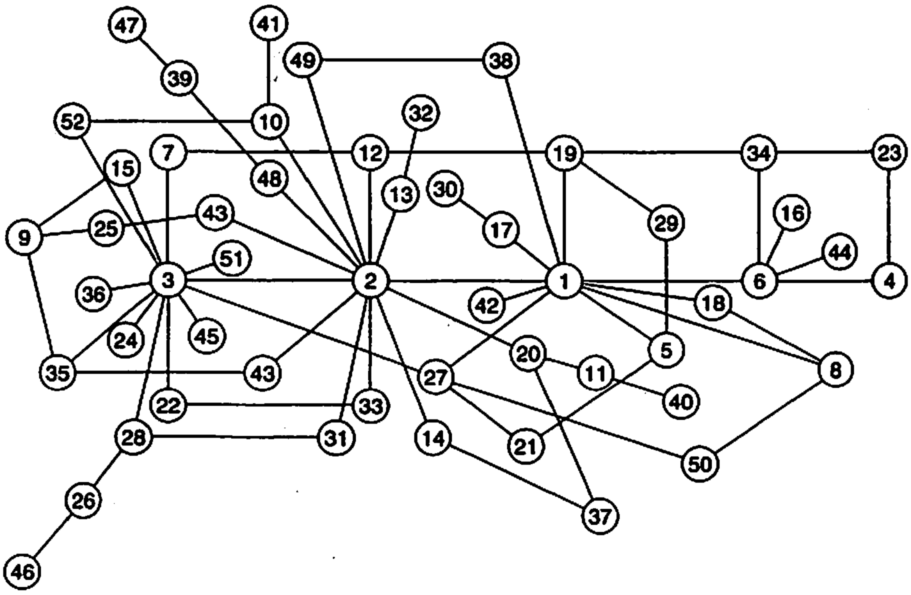

Fig. 6. A phylogenetic network for 52 human mtDNA sequences.

References

- Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990) Basic local alignment search tool. Journal of Molecular Biology, vol. 215, pp. 403-410.

- Bandelt H.-J., Forster P., Sykes B. C., and Richards M. B. (1995) Mitochondrial portraits of human populations using median networks. Genetics, vol. 141, pp. 743-753.

- Oota H., Saitou N., Matsushita T., and Ueda S. (1999) Molecular genetic analysis of a 2,000-year old human population in China. American Journal of Human Genetics, vol. 64, no. 1, pp. 250-258.

- Saitou N. and Yamamoto F. (1997) Evolution of primate ABO blood group genes and their homologous genes. Molecular Biology and Evolution, Vol. 14, No. 4, pp. 399-411.